Drive&Act: Driver Action Recognition

Presented By:

Purisa Jasmine Simmons, Gates Zeng, Yuting Jiang, Andrew Pang

Mentors: Professor Trivedi, Akshay, Bowen, Ross
Objective

- To classify in-cabin driver’s activities using Drive & Act dataset based on Inflated 3D ConvNets (I3D), specifically: Improve the performance of the I3D model beyond current baseline on fine-grained activities
Methodology

- The paper's best-performing model on end-to-end fine-grained activity recognition was I3D (compared to C3D and P3D ResNet) so we decided to explore the I3D model further
- In the paper, the authors separately train models using kinect IR, Color, and Depth data and softmax average over all cameras
- In our paper, we would like to train a single I3D model with 4 streams, one for each kinect IR, Color, Data, and an additional FLOW stream
- We also use transfer learning by using weights pre-trained on ImageNet
Methodology Flowchart

- Literature Review
- Data Collection & Pre-Processing
- Implement Model Architecture
- Reproduce Baseline with each sensor separately
- Combine RGB, FLOW for baseline 2-Stream-I3D
- Combine RGB, FLOW, Depth, IR for 4-Stream-I3D
- Final Report
Expected Goals

- We expect our model to outperform the original Drive&Act Dataset baseline since we include an additional FLOW stream, giving the model additional helpful information to classify drivers' actions.
- We also use transfer learning by using weights pre-trained on ImageNet, which the Drive&Act authors do not use, but which was shown to be helpful in improving accuracy in the original I3D paper.
Significance and Novelty

• Original I3D authors propose a 2-stream model for RGB and FLOW data, we have yet to see a 4-stream model
• Not much work has been done on this dataset yet other than the benchmark proposed by the original authors
• Creating a harder benchmark could encourage others to use this dataset and improve efforts which focus on creating a higher standard of driver behavior analysis
• Overall, this could significantly improve safety around drivers in AV
Published Oct. 2019
Created publicly available dataset to facilitate research of activity recognition under realistic driving conditions
Provides challenging benchmarks for video and body-based activity recognition

Driver secondary activities in context of both autonomous and manual driving (83 classes in total).
Multi-modality: color-, depth-, infrared- and body pose data, as conventional RGB-based action recognition datasets disregard the case of low illumination.
Multi-view: six synchronized camera views cover the vehicle cabin to deal with limited body visibility.
Hierarchical activity labels on three levels of abstraction and complexity, including context annotations.
Fine-grained distinction between individual classes (e.g. opening bottle and closing bottle) and high diversity of action duration and complexity, which poses an additional challenge for action recognition approaches (e.g. opening door from inside often takes less than a second while reading a magazine might last for minutes).
Inflated 3D ConvNet (I3D): Action recognition model proposed by Carreira, Zisserman
Current Progress

- Wrote scripts to extract data frames from dataset, preprocess and resize videos, and label frames with correct activities
Current Progress

- Training and testing I3D models against each dataset.
  - Pretraining with Imagenet model

- Working on getting acceptable baseline results for I3D model run on each of the subsets of data that we have from each sensor: Kinect IR, RGB, Depth, FLOW

- Running into issues with our I3D model having very low accuracies - working with Akshay to resolve these issues
I3D Modifications:

- Localization Loss, Classification Loss
  - Our data only had one activity for each video segment → Localization Loss is not necessary
  - Only needed classification loss

- Modified classification task
  - Rather than taking the most frequent predicted activity in the frame sequence
  - Task was predicted by the last frame instead
  - Last frame contained the largest receptive field, encapsulating information from all previous frames
Our Results

- Currently using a subset of the provided dataset. (~20%) Due to storage limitations.
### Table 2: Fine-grained Activities recognition on the Drive&Act validation and test set. We group our proposed models into: (1) baselines, (2) networks that only use the body pose representation and (3) CNN-based end-to-end methods that make predictions directly on the input images.

<table>
<thead>
<tr>
<th>Type</th>
<th>Model</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Random</td>
<td>2.94</td>
<td>2.94</td>
</tr>
<tr>
<td>Pose</td>
<td>Interior</td>
<td>45.23</td>
<td>40.30</td>
</tr>
<tr>
<td></td>
<td>Pose</td>
<td>53.17</td>
<td>44.36</td>
</tr>
<tr>
<td></td>
<td>Two-Stream</td>
<td>53.76</td>
<td>45.39</td>
</tr>
<tr>
<td></td>
<td>Three-Stream</td>
<td>55.67</td>
<td>46.95</td>
</tr>
<tr>
<td>End-to-end</td>
<td>C3D</td>
<td>49.54</td>
<td>43.41</td>
</tr>
<tr>
<td></td>
<td>P3D ResNet</td>
<td>55.04</td>
<td>45.32</td>
</tr>
<tr>
<td></td>
<td>I3D Net</td>
<td><strong>69.57</strong></td>
<td><strong>63.64</strong></td>
</tr>
</tbody>
</table>

### Table 5: Fine-grained activity level results for different views and modalities and their combination (I3D model).

<table>
<thead>
<tr>
<th>Camera</th>
<th>View</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIR</td>
<td>front top</td>
<td>69.57</td>
<td>63.64</td>
</tr>
<tr>
<td></td>
<td>right top</td>
<td>65.16</td>
<td>60.80</td>
</tr>
<tr>
<td></td>
<td>back</td>
<td>54.70</td>
<td>54.34</td>
</tr>
<tr>
<td>Cameras</td>
<td>face view</td>
<td>49.73</td>
<td>42.98</td>
</tr>
<tr>
<td></td>
<td>left top</td>
<td>68.72</td>
<td>62.83</td>
</tr>
<tr>
<td></td>
<td>combined</td>
<td>72.70</td>
<td>67.17</td>
</tr>
<tr>
<td>Kinect Color</td>
<td>right top</td>
<td>69.50</td>
<td>62.95</td>
</tr>
<tr>
<td>Kinect Depth</td>
<td></td>
<td>69.43</td>
<td>60.52</td>
</tr>
<tr>
<td>Kinect IR</td>
<td></td>
<td>72.90</td>
<td>64.98</td>
</tr>
<tr>
<td>Combined</td>
<td></td>
<td>73.80</td>
<td>68.51</td>
</tr>
<tr>
<td>All combined</td>
<td></td>
<td><strong>74.85</strong></td>
<td><strong>69.03</strong></td>
</tr>
</tbody>
</table>
Breakdown of Individual Workload

Gates
- Kinect Color RGB
- I3D Test Script
- Model Training

Jasmine
- Kinect IR
- Pre-processing input data
- I3D Model Setup
- Model Training

Andrew
- Kinect IR (Optical Flow)
- Pre-Processing Input Data
- Model Training

Yuting
- Kinect Depth
- Pre-Processing Input Data
- Model Training
Future Tasks

1. Improve I3D model accuracy on kinect_ir and kinect_color data
   a. Hopefully, obtain space for more data to be trained with.

2. Implement Two-Stream approach for our I3D model.

3. Augment I3D model to accept 4 streams of data
   a. The model currently accepts a single 3-channel RGB stream and a 2-channel image flow stream
   b. We want to add two more 3-channel streams to the model (Kinect IR, Kinect Depth)

4. Evaluate the modified I3D model and analyze results

5. Write final report
Project Plan Timeline

Week 10
1) Improve the accuracy of each single stream model to get close to the results in the Drive&Act paper
2) Modify model to 2 stream i3D model with Kinect RGB and optical flow

Week 11
1) Modify the model to add in 2 additional streams for a total of 4 streams with Kinect RGB, FLOW, Depth, IR
2) Write final report
Thank you for your attention!

Q&A